INTRODUCTION TO READ MAPPING

Dhivya Arasappan

(With several slides borrowed from Dr. Jeff Barrick)

What does an alignment look like?
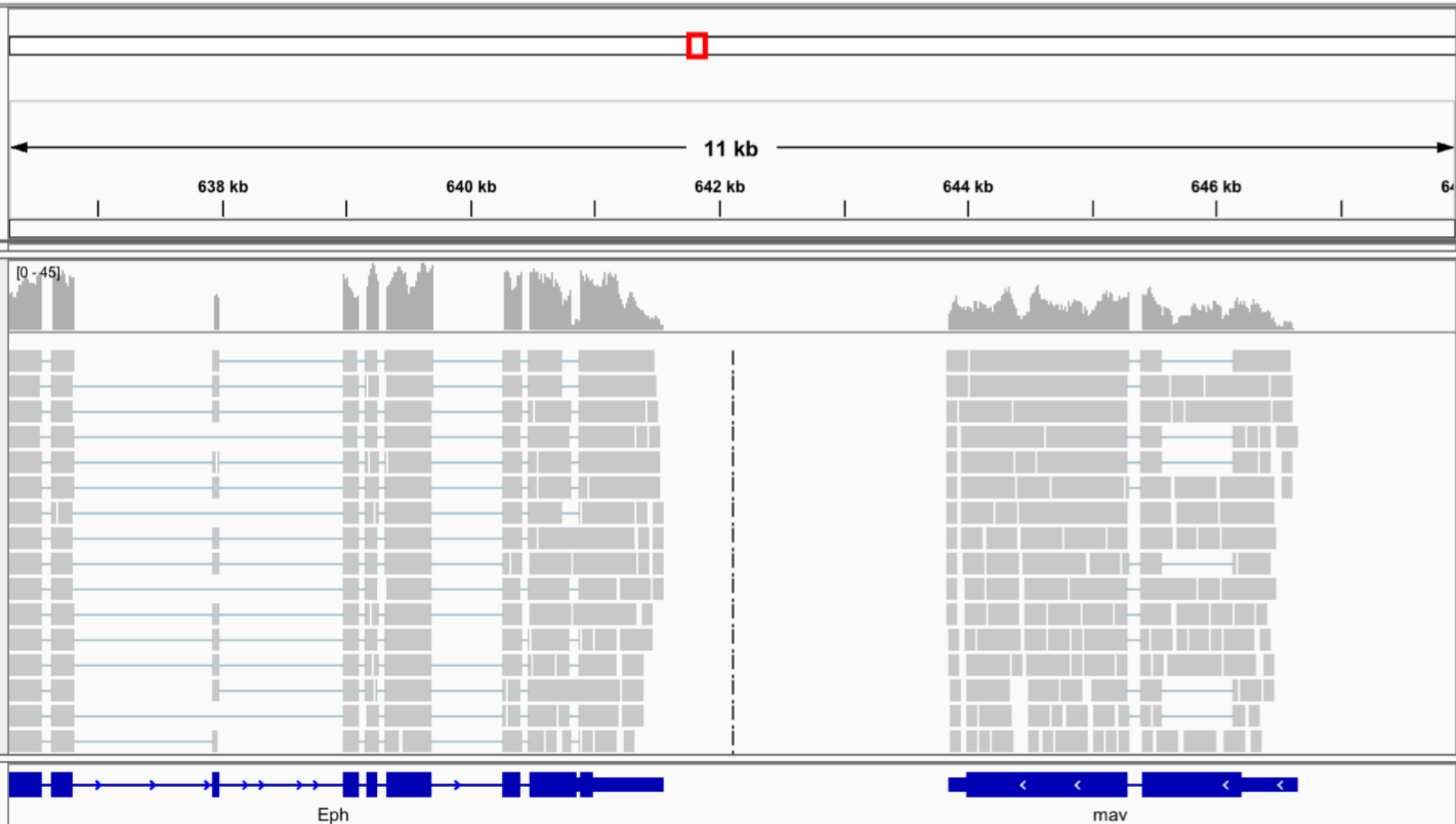
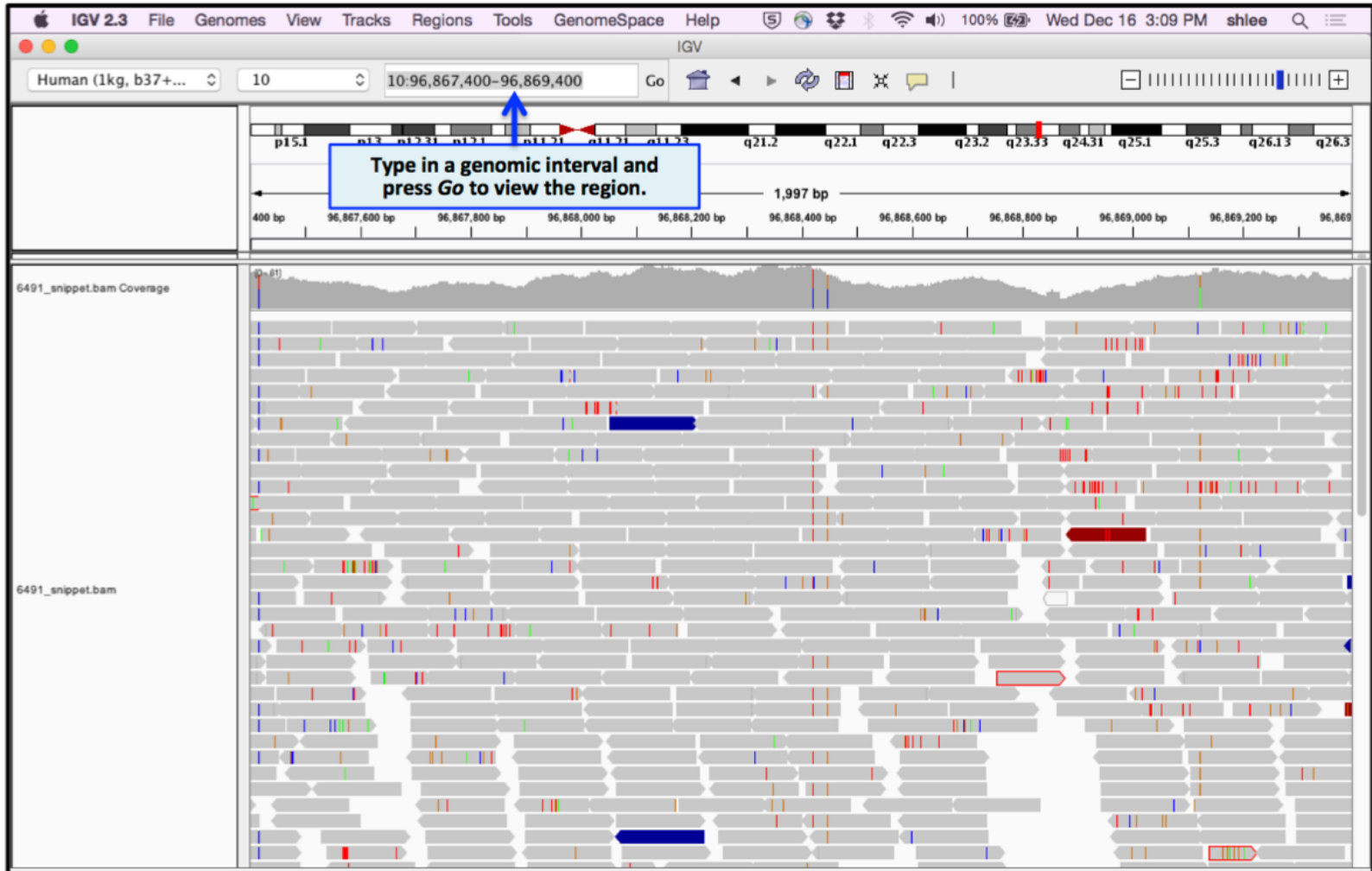Ref=TAGATCAGATTCGATACCAGACCATGATCATACGATCCA

Read=AGACCATG

Found at offset 18!
TAGATCAGATTCGATACC**AGACCATG**ATCATACGATCCA

What does an alignment look like?

What is an alignment look like?

Why is alignment a difficult problem?

- 100's of millions of reads

- Billions of bases to search through

- Approximate matching

- Looking for a tiny pattern (~100-120 bp read) in a large, often redundant sequence.

Basic steps of mapping reads

- Pre-mapping QC
- Build a reference sequence index.
- Map sequencing reads to the reference index.
- Convert results to SAM/BAM format and obtain mapping statistics.
- Post-mapping analysis.

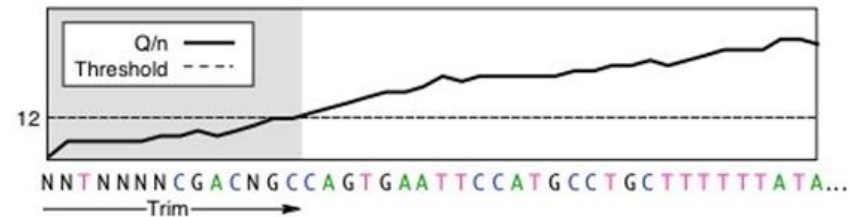# What to know about your data before mapping?

KNOW YOUR DATA!

- Paired end? Single end?
- Traditional RNA-Seq? 3' tag
- Insert size estimate?

PREPROCESSING

- Adaptor sequences trimmed?
- Primer sequences/barcodes removed?
- Poor quality regions trimmed?

What will your reads look like?

FASTQ FORMAT

@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
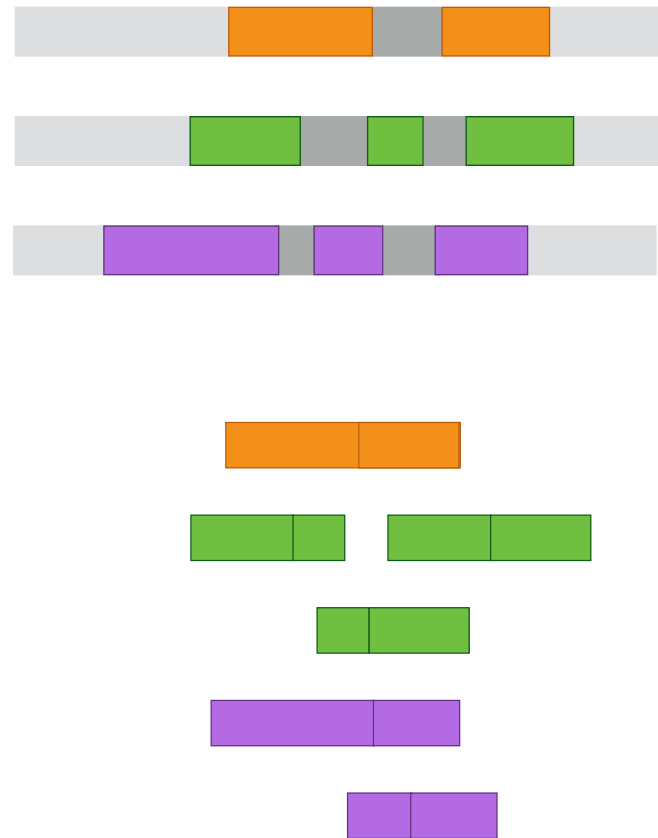B@BBBBB@BBBBAAAA>@AABA?BBBAAB??>A?

Line 1: @read name
Line 2: called base sequence
Line 3: +read name (optional after +)
Line 4: base quality scores

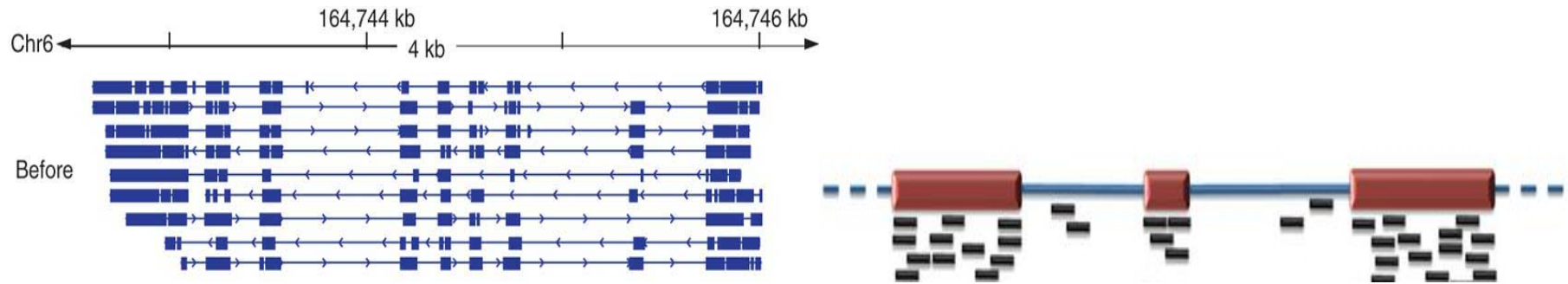# What does the reference look like?

- **Genome:** All the DNA of an individual, organized by chromosome, containing non-coding and coding regions.

- **Transcriptome**: All the gene isoforms. No non-coding sequences.

What to know about your reference before mapping?

- Mapping to genome vs transcriptome?

- Is your reference the right version?
- Does your annotation match your reference?

What will your reference look like?

- FASTA Format

  >gi|254160123|ref|NC_012967.1| Escherichia coli B str. REL606
  agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
  tgatagcagcttctgaactggttacctgccgtgagtaaattaaaatttattgacttagg

tcactaaatactttaaccaatataggcatagcgcacagacagataaaaattacagagtac

acaacatccatgaaacgcattagcaccaccattaccaccaccatcaccattaccacaggt
  ….

- Using complex reference sequence names is a common problem during analysis. Might rename:

  >REL606
  agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
  tgatagcagcttctgaactggttacctgccgtgagtaaattaaaatttattgacttagg

What will your annotation look like?

- GFF3 Format
  - seqname - The name of the sequence.
  - source - The program that generated this feature.
  - feature – Examples: "CDS", "start_codon", "stop_codon", and "exon".
  - start - The starting position of the feature in the sequence.
  - end - The ending position of the feature (inclusive).
  - score - A score between 0 and 1000.
  - strand - Valid entries include '+', '-', or '.' (for don't know/don't care).
  - Frame – reading frame
  - group – ID and other information about the entry

  Example:
  Rel606   refseq   cds 1450   1540   500 +   .   Gene_id=« test_gene »

- Make sure the GFF3 file matches your reference fasta file.

Where to get your references?

- Ensembl ftp
- UCSC
- Gencode
- Organism specific databases/websites.

First Step : Reference Indexing

- **Indexing**: Think of the index of the book.

- Break reference into substrings of K length (Kmers) and index all locations of the Kmer.

Reference: TTACTTTACG

ACG	6
ACT	3
CTT	4
TAC	2,7
TTA	1,6
TTT	5

Read: CTTAAC

Find the prefix and extend the alignment

Indexing

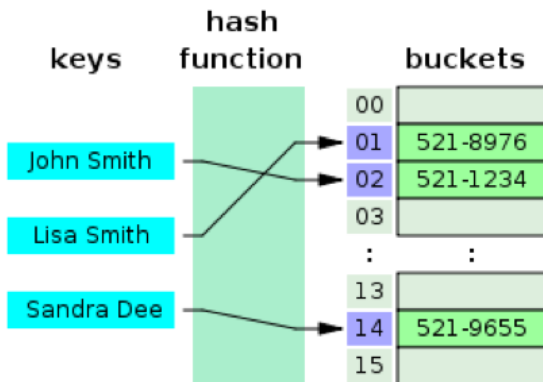- Many different ways to represent reference indexes

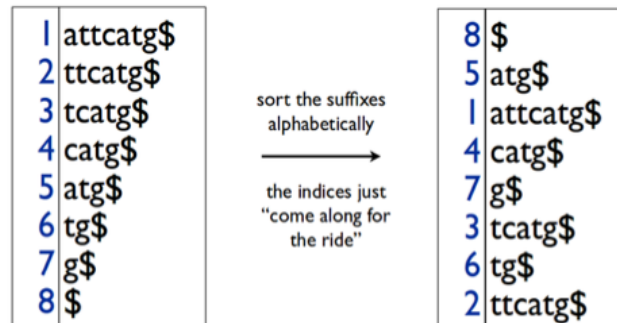## hash table

image from wikipedia

suffix array

image from discuss.codechef.com
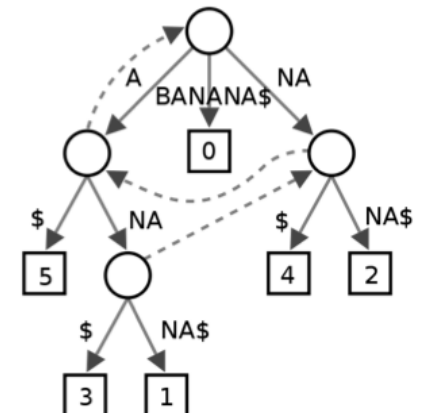
suffix tree

image from wikipedia

Types of Mappers

Class	Category	Package	Notes	Uses
Read mapping				
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome
		Stampy[39]	Probabilistic model	
	Burrows-Wheeler transform methods	Bowtie[43]		
		BWA[44]	Incorporates quality scores	
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions
		SpliceMap[50]		
		TopHat[51]	Uses Bowtie alignments	
	Seed-extend methods	GSNAP[53]	Can use SNP databases	
		QPALMA[54]	Smith-Waterman for large gaps	

Figure :
Garber et al, Nature Methods, 2011

Unspliced Mapping
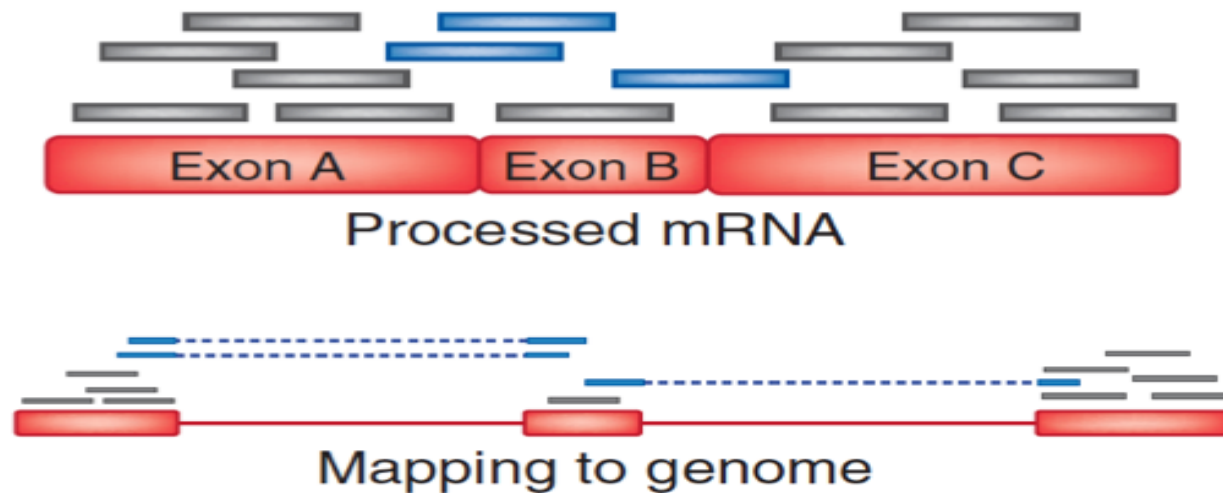
Class	Category	Package	Notes
Read mapping			
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension
		Stampy[39]	Probabilistic model
	Burrows-Wheeler transform methods	Bowtie[43]	
		BWA[44]	Incorporates quality scores

Figure :
Garber et al, Nature Methods, 2011

b   Burrows-Wheeler

Reference genome (> 3 gigabases)   Short read

Chr1
Chr2
Chr3
Chr4

ACTCCCGTACTCTAAT

Concatenate into single string

Burrows-Wheeler transform and indexing

Bowtie index (~2 gigabytes)

ACTCCCGTACTCTAAT

T
AT
AAT

Look up 'suffixes' of read

ACTCCCGTACTCTAAT

Hits identify positions in genome where read is found

Convert each hit back to genome location
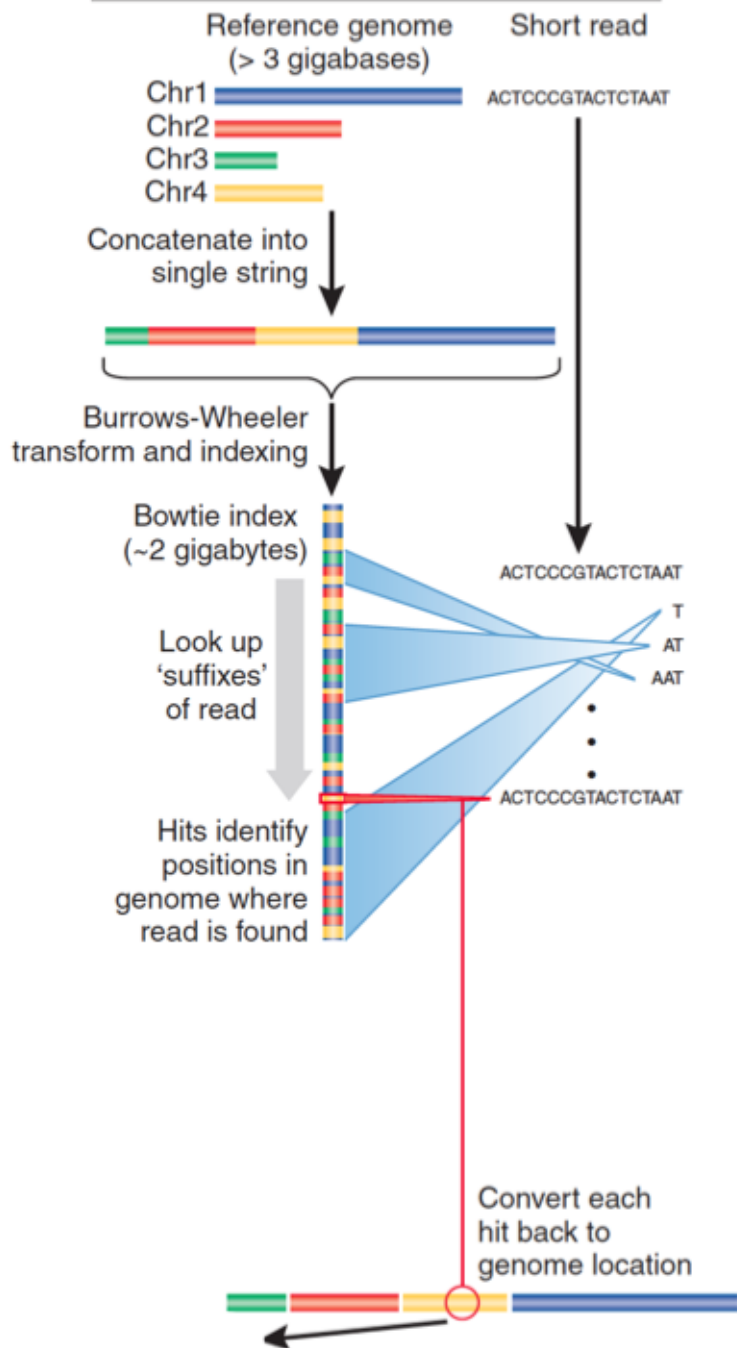
Burrows-Wheeler transform compresses sequence.

Input	`SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES`
Output	`TEXYDST.E.IXIXIXXSSMPPS.B..E.S.EUSFXDIIOIIIT`

## Suffix tree enables fast lookup of subsequences.
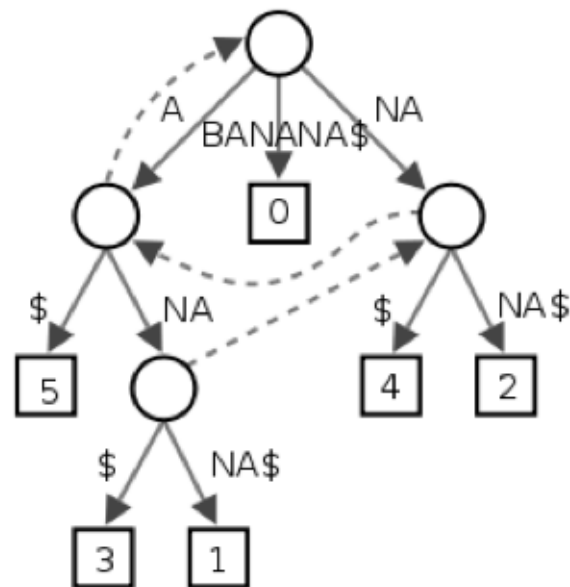
http://en.wikipedia.org/wiki/Suffix_tree

Exact matches at all positions below a node.

Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).
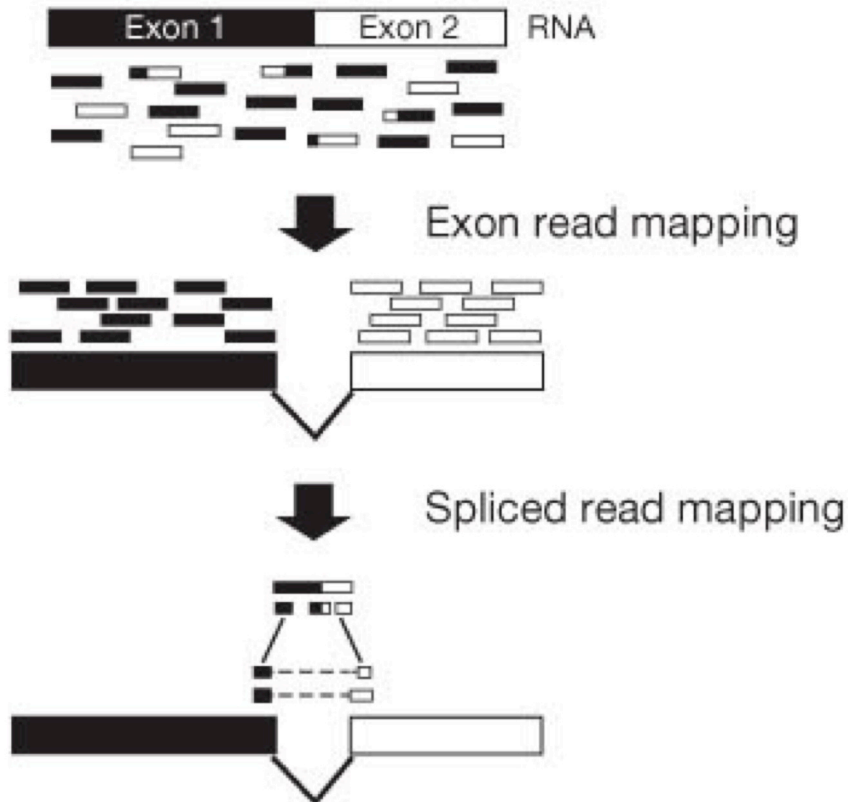
Spliced mapping

- Needed for quantifying and identifying splice variants from RNA Seq data.

- Tools:
  - HISAT2
  - STAR
  - Tophat
  - SpliceMap
  - MapSplice
  - RUM
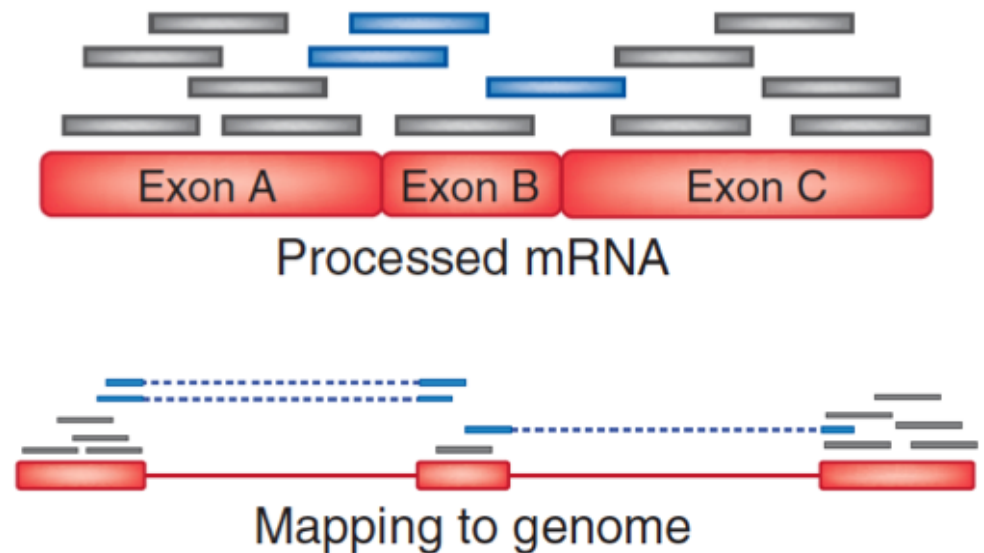
Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).
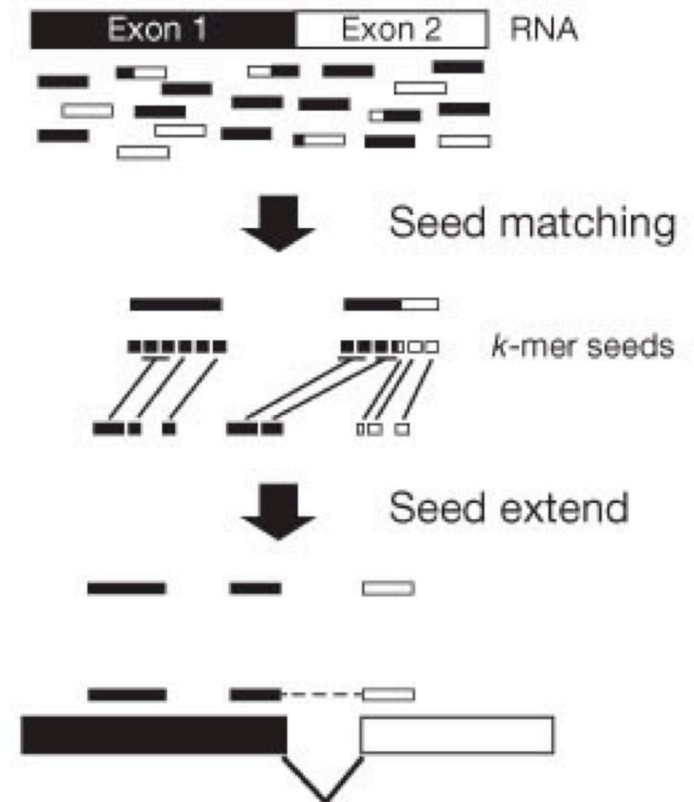
Spliced mapping

Figure :
Garber et al, Nature Methods, 2011

Paired end mapping

paired-end

two inwardly oriented
reads separated by ~200 nt

- Pairs map with expected insert size.
- One part of the pair, after mapping, is the anchor for the next read's mapping.

PAIRED END READ MAPPING IS VERY HELPFUL IN RNA SEQ!!

gene X   gene Y

http://genomebiology.com/content/12/1/R6
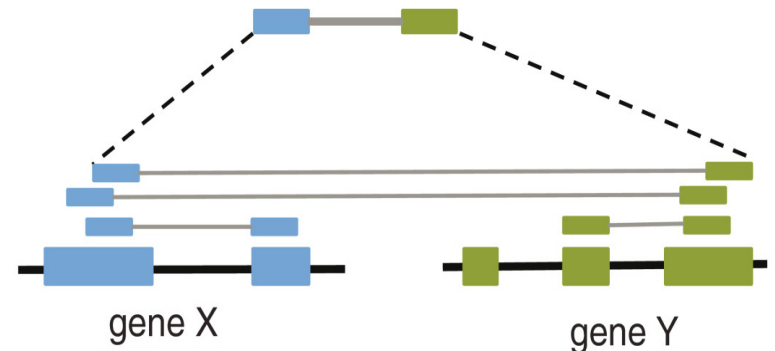
Mapping Quality

- Mapping quality is the probability that a read is aligned to the wrong place.

$$p = 10^{**}(-q/10)$$

- BWA mapping quality calculated by considering:
    - Repeat structure of reference
    - Read alignment quality (mismatches etc)
    - Number of mappings
    - BWA will assign a mapping quality of 0 to reads that mapped equally well to multiple places

Mappers comparisons

Accuracy Performance of Aligners

Legend:
- Correctly Mapped
- Incorrectly Mapped
- Ambiguously Mapped
- Unmapped

x-axis: Fraction of total reads

y-axis categories: bwa, gsnap, rum, star, tophat
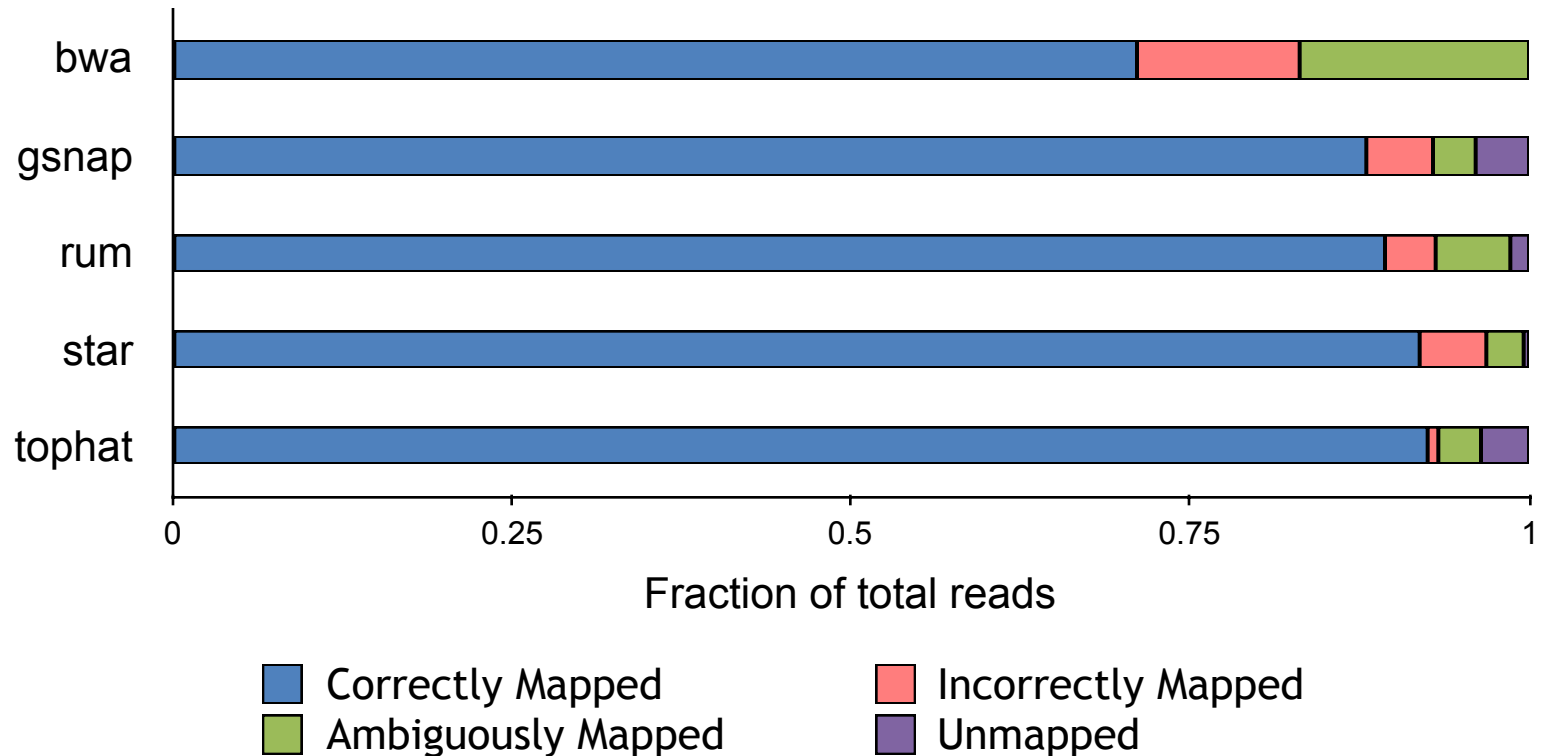
New benchmarking analysis
performed by **Raghav Shroff**

Mappers comparison

CPU time for RNA-seq Aligners
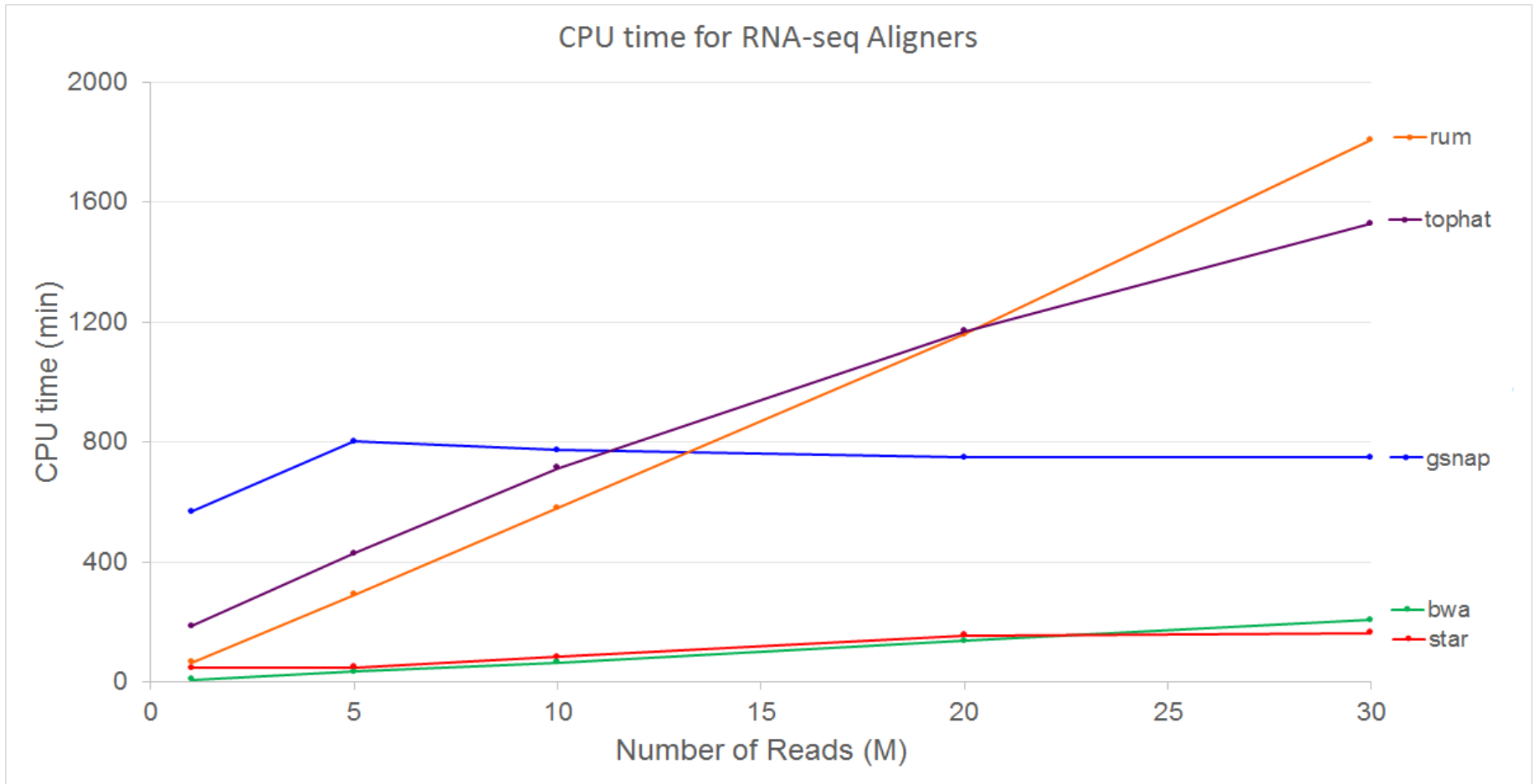
New benchmarking analysis
performed by **Raghav Shroff**

SAM file format

- Alignment results generated in Sequence Alignment/Map format

- Tab delimited, with fixed columns followed by user-extendable key:data values.

- Most mappers also output unmapped reads in SAM file.

- SAMTOOLS – toolkit to manipulate, parse SAM files.

SAM File Format

SAM fixed fields:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2^{29}-1]	Position of the mate/next segment
9	TLEN	Int	[-2^{29}+1,2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

```
SRR030257.264529      99  NC_012967  1521     29  34M2S   =     1564
      79  CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
      AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
      XT:A:M  NM:i:3  SM:i:29  AM:i:29  XM:i:3  XO:i:0  XG:i:0  MD:Z:23T0G4T4
```

CIGAR score

```
Ref    CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC!
Read          aaATGTCGCGGTG.TAGGAggatcc!
```

2S5M2I4M1D4M6S

	Op	BAM	Description
	M	0	alignment match (can be a sequence match or mismatch)
	I	1	insertion to the reference
	D	2	deletion from the reference
*	N	3	skipped region from the reference
	S	4	soft clipping (clipped sequences present in SEQ)
*	H	5	hard clipping (clipped sequences NOT present in SEQ)
*	P	6	padding (silent deletion from padded reference)
*	=	7	sequence match
*	X	8	sequence mismatch

*Rarer / newer

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

BAM format

- SAM files are converted to BAM format through SAMTOOLS command:

    - samtools view –b –S samfile > bamfile

- BAM file is binary format.

- BAM file is compressed.

- BAM files are usually what you need for post mapping analysis and visualization.

TAKEAWAYS

- Unspliced mapper-
  - Most suited for mapping to transcriptome
  - Example: BWA

- Spliced mapper-
  - Most suited for mapping to genome
  - Example: Hisat2, Star

- Mapping output
  - SAM File: tab-delimited file
  - Filter SAM file, Assess mapping stats

SAM FILE FLAGS EXPLAINED

QNAME   SRR035022.2621862
FLAG   163

The QNAME is the query name. For the FLAG of 163 we transform this into a binary string: 10100011. So accordingly to the flag table:

Flag	Description
0x0001	the read is paired in sequencing, no matter whether it is mapped in a pair
0x0002	the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment) [1]
0x0004	the query sequence itself is unmapped
0x0008	the mate is unmapped [1]
0x0010	strand of the query (0 for forward; 1 for reverse strand)
0x0020	strand of the mate [1]
0x0040	the read is the first read in a pair [1,2]
0x0080	the read is the second read in a pair [1,2]
0x0100	the alignment is not primary (a read having split hits may have multiple primary alignment records)
0x0200	the read fails platform/vendor quality checks
0x0400	the read is either a PCR duplicate or an optical duplicate

1   the read is paired in sequencing, no matter whether it is mapped in a pair
1   the read is mapped in a proper pair
0   not unmapped

0   mate is not unmapped
0   forward strand
1   mate strand is negative
0   the read is not the first read in a pair
1   the read is the second read in a pair